Chronicling America, National Digital Newspaper Program, HDNP: Technical Aspects

Challenges of Newsprint

- Newspapers are a difficult medium
- Never meant to last, made for daily use and disposal
- Pages crumble and acid corrodes the materials
- Tracking serial publications over time
- Patron demand increased, storage space grew scarce, binding costs rose

Microfilm

- Adopted in the 1920s as a standard
- Turned newspaper from a storage nightmare to a relatively easy medium to handle
- Libraries had to decide what to do with the hardcopy
  - Keep in holdings?
  - Deaccession?

USNP 1980-2007

- Funded by National Endowment for the Humanities, managed by the Library of Congress
- Goals: Locate, catalog, and microfilm newspapers
    - Created bibliographic records for over 140,000 newspaper titles; provided access to 70 million pages of newsprint in microfilm
- University of Hawai'i with Hawaiian Historical Society, Hawai'i State Archives and State Library contributed for Hawai'i
- Hawai'i microfilmed 260,000 pages and cataloged 476 titles

NDNP

- Enhance access to newspapers, build on foundation of the USNP by creating a national resource of historically significant newspapers from all the states and U.S. territories
- Establish technical conversion specs & practices for efficient basic discovery & access
- Develop production tools to ensure good digital objects that can be managed & preserved long-term
- Take preservation responsibility for the digitized newspapers

NDNP Program

- Began in 2005 with 6 state participants
- 2-Year awards to state projects, renewable
- Each state/phase to digitize 100,000 pages of microfilmed newspaper
- Newspapers picked must be from between 1836 and 1922
- Participants must write historical essays on each newspaper

Challenges

- Where are the master reels?
- Copyright issues (Who filmed the newspapers & owns the master microfilm)
- Technical specifications (Poorly filmed, low density readings)
- Microfilm standards applied vary widely
- No universally accepted metadata standard for historical newspapers
- Titles, issues, pages and reels all need to be represented as different yet related classes of objects

TECHNICAL SPECIFICATIONS - DELIVERABLES

- Images scanned at 300-400 dpi in three image formats:
- Grayscale, uncompressed Tiff 6.0
- Compressed JPEG2000
- PDF Image with hidden text
  - digital formats with a high probability of sustainability
- ***Provide structural and technical metadata***
- OCR text for all pages
- Capture grayscale preservation microfilm targets

Deliver all digital assets in METS object structure

METS (Metadata Encoding and Transmission Standard) includes metadata used to relate pages to title, date, and edition; sequence pages within issue or section; and to identify image and OCR files

XML Schema for creating XML files that define the hierarchical structure of digital library objects (images, text files, etc.); the names and locations of the files and associated metadata (e.g., MODS)

DOWN THE RABBIT HOLE OF ACRONYMS ...

METS – Metadata Encoding and Transmission Standard

MODS – Metadata Object Description Schema

PREMIS – PREservation Metadata Implementation Strategies

Sections of a METS file

<mets>

<metsHdr/> *METS header (document talks about itself)*

<dmdSec/> *Descriptive metadata (MODS, etc.)*

<amdSec/> *Administrative metadata (copyright info., etc.)*

<fileSec/> *File section (names and locations of files)*

<structMap/> *Structural map (relationships of the parts)*

<structLink/> *Linking information*

<behaviorSec/> *Binding executables/actions to object*

</mets>

OCR requirements

- Conform to ALTO XML schema
  ALTO *(Analyzed Layout and Text Object) is a XML (Extensible Markup Language) Schema that details technical metadata for describing the layout and content of physical text resources*
- Provide bounding box coordinate data
  Each column is sectioned and coordinates are used to place words

OCR/OWR

- Optical Character Recognition/Optical Word Recognition does not yield article “transcriptions”
- Text OCR’d from images of newspapers is used for searching purposes
- The bounding box coordinate information allows for several search options
  - ANY of the words, ALL of the words, exact PHRASE
  - Proximity search
  - Look for words within 5, 10, 50 words of one another

WHY?

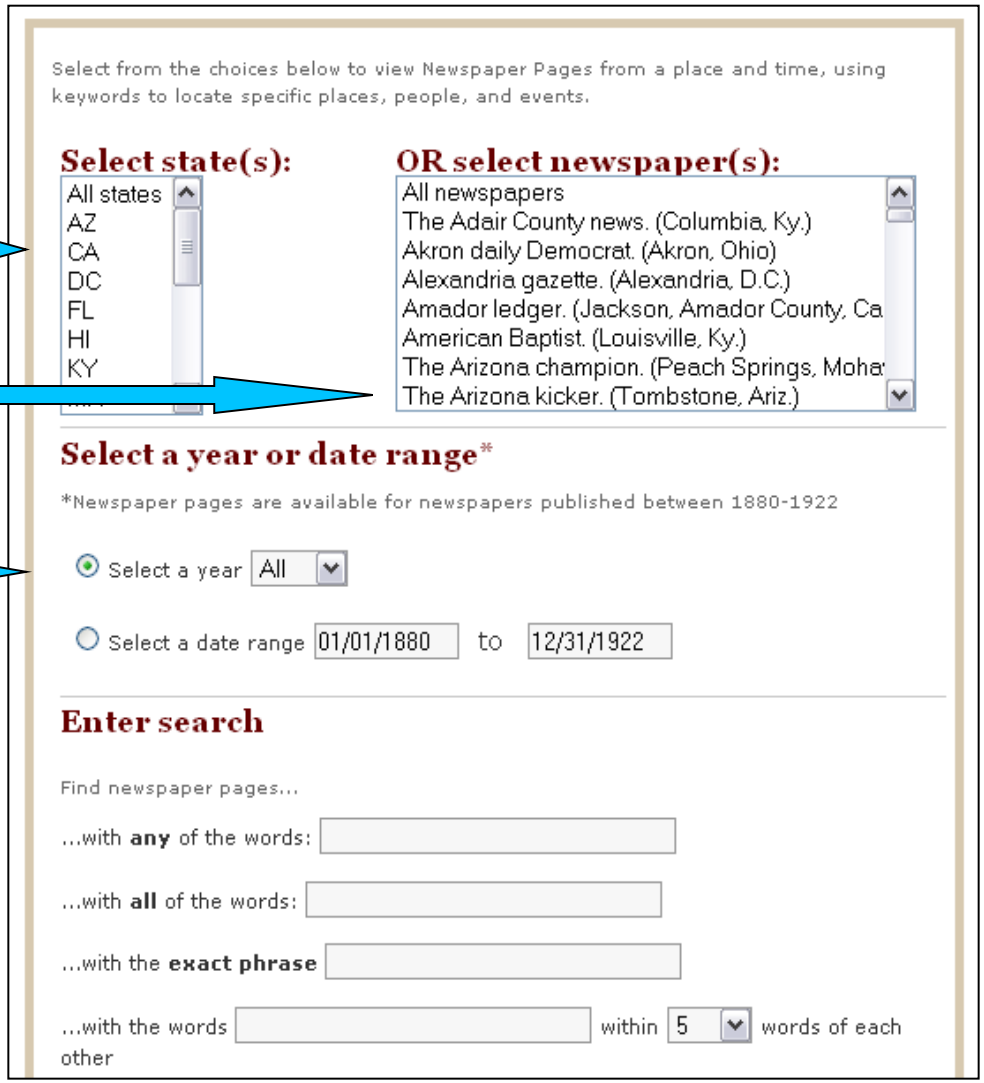
- XML structure is used by software for generating/creating multiple outputs:
  - HTML/XHTML for Web display
  - PDF for printing
- Ease of automated editing (single records or batches of records)
- Ability to validate data
- Interoperability (e.g. Repository submission and OAI harvesting
- Create and improve User Interface without having to redo underlying source files

All that coding pays off for the user when SEARCHING

Geographic metadata

Title metadata

Date metadata



Metadata allows interface to offer display options

- Click on thumbnail or description of page to view larger version
- Go to Next, Jump to page ...
- Change to List View or back to Thumbnail view

Your fulltext search returned 97855 results

Display: List View | Thumbnail View

Sort by: Relevance | State | Title | Date

Previous | 1 2 3 ... 9784 9785 9786 | Next Jump to page: 1

Austin's Hawaiian weekly. (Honolulu [Hawaii) 1899-190?, June 17, 1899,

Austin's Hawaiian weekly. (Honolulu [Hawaii) 1899-190?, June 17, 1899,

Metadata is used to generate a calendar view

- Visually indicate which issues have been digitized
- Change which year you're viewing
- Browse First Pages

About this Newspaper | Libraries that Have It | Browse Issues | MARC Record | More About this Newspaper

Browse Calendar | Browse First Pages

Single edition = dates in **bold**.
Multiple editions = dates in ***bold italics***.

Issues for 1899

January, 1899

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

February, 1899

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28				

March, 1899

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

April, 1899

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

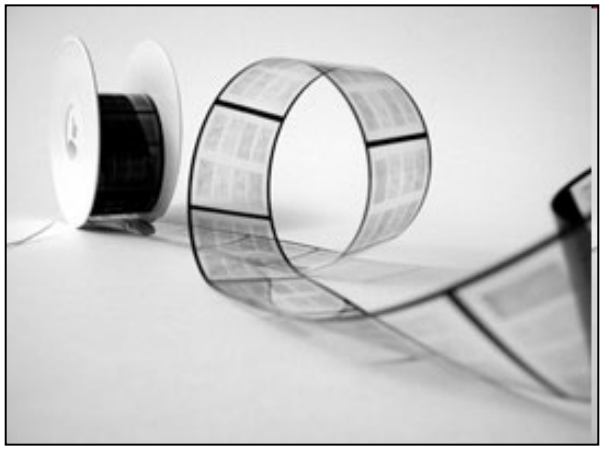
May, 1899

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

June, 1899

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

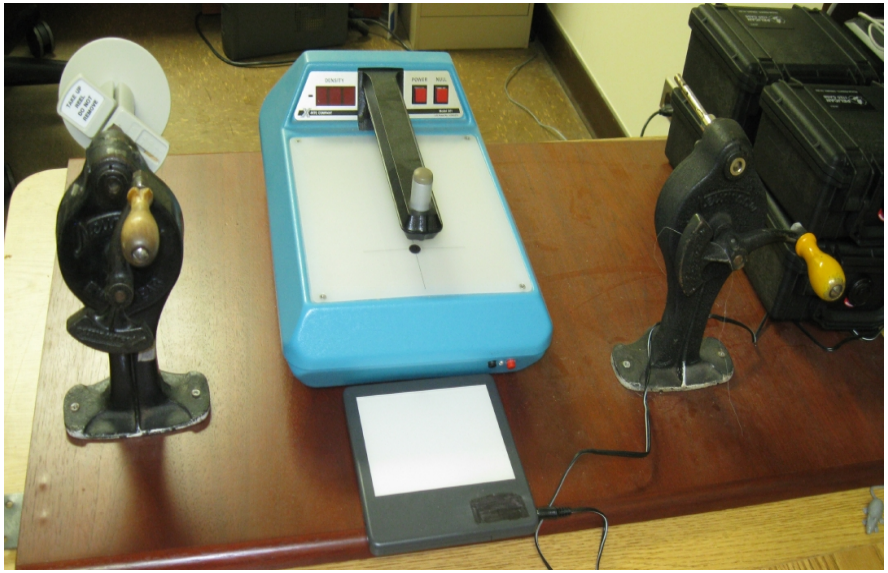
From Microfilm to Digital Images



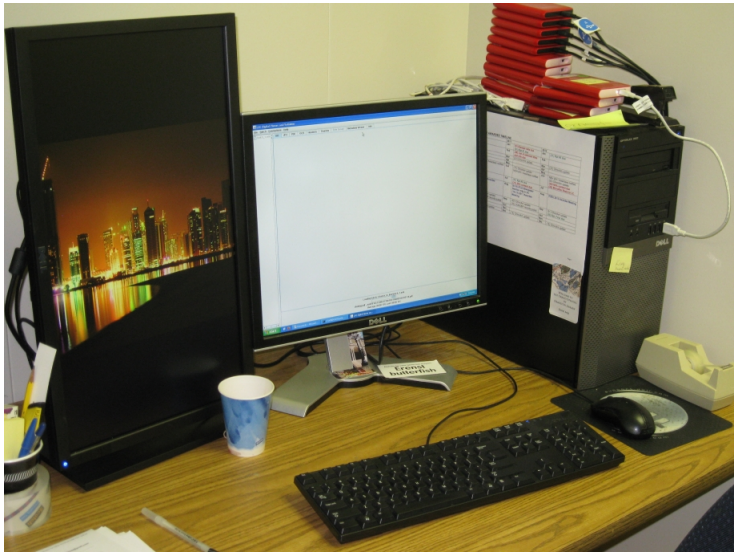
- Request for Proposals (RFP) include all LC technical specifications
- Position Description(s) Coordinator, students
- Hiring and Training

- Title Selection (copyright checks)
  - Microfilm identification & duplication
- Evaluate RFPs, select Vendor
- Digitization
- Metadata creation & Validation

- External Hard Drives and Pelican cases
- 1 PC with double monitor
- Densitometer
- Microfilm reader/scanner
- Library of Congress' Digital Validator & Viewer (DVV)



Some of our stuff

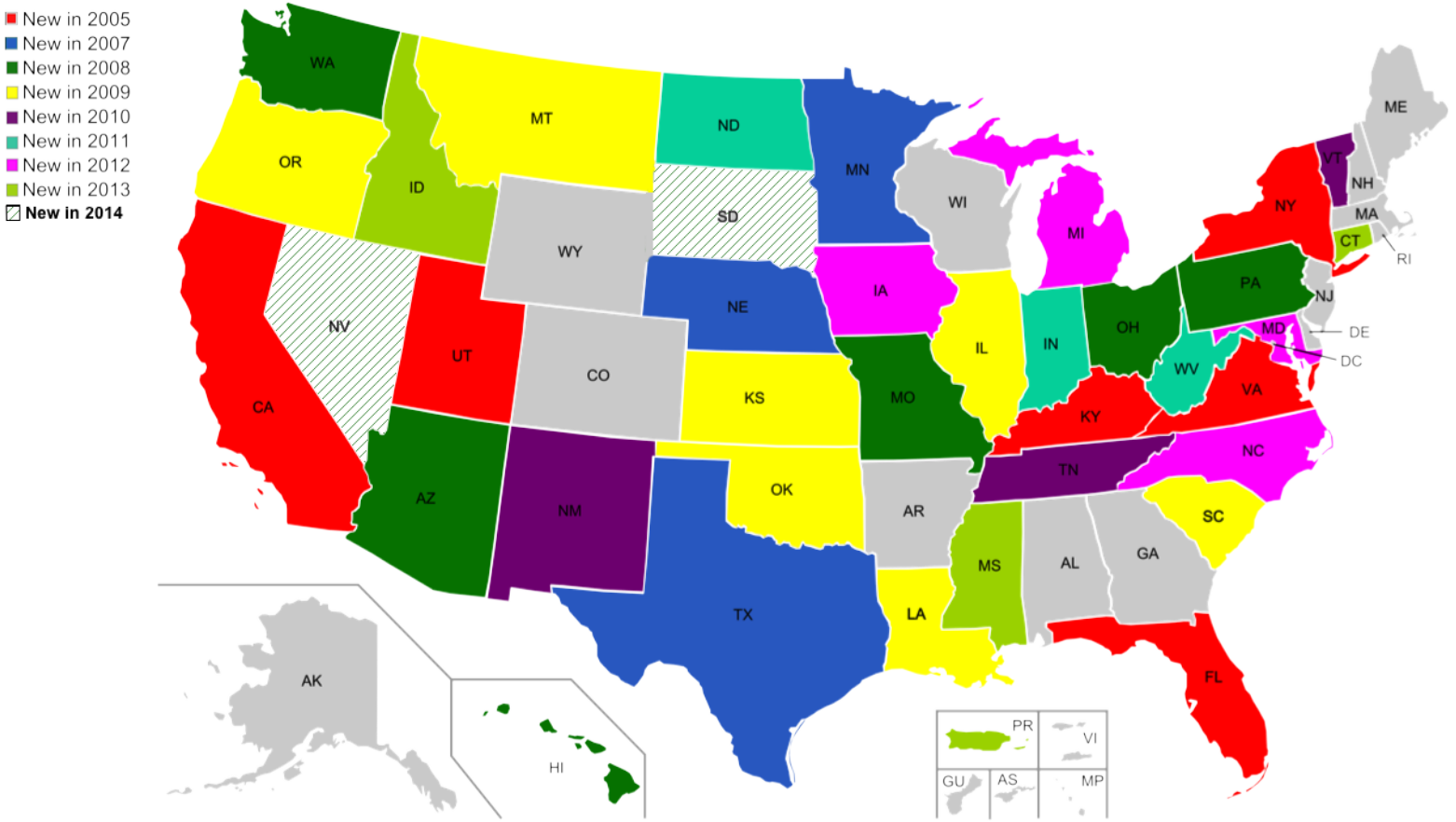


QUALITY CONTROL

Image quality: Too dark? Too light? Skewed?
Correct image? Compare digitized image to microfilmed image

Check for Missing Issue/Page tags
Review metadata

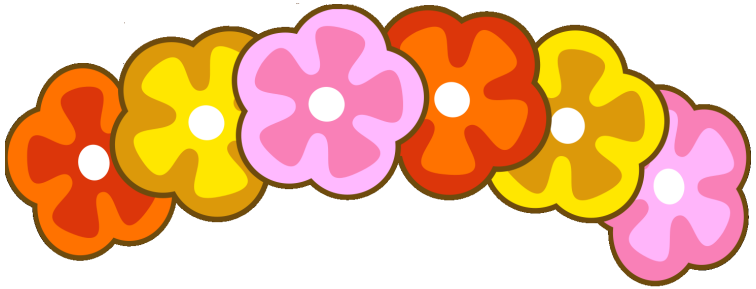
- Dates
- LCCN #
- Locations



37 states, Puerto Rico & District of Columbia

LINKS

- Chronicling America: http://chroniclingamerica.loc.gov/
- Library of Congress: http://www.loc.gov/ndnp/
- National Endowment for the Humanities: http://www.neh.gov/projects/ndnp.html
- Hawai'i Newspapers: a union list http://evols.library.manoa.hawaii.edu/handle/10524/2089
- Using <METS> and <MODS> to Create XML Standards-based Digital Library Applications http://www.loc.gov/standards/mods/presentations/mets-mods-morgan-ala07/



Thank you – Mahalo

Now on to ...